dGAMLSS: An exact, distributed algorithm to fit Generalized Additive Models for Location, Scale, and Shape for privacy-preserving biomedical inference

Fengling Hu1,✉, Jiayi Tong2, Yong Chen2, and Russell T. Shinohara1,3

24 May, 2024

1 Penn Statistics in Imaging and Visualization Endeavor (PennSIVE), Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania  
2 Penn Computing, Inference, and Learning Lab (PennCIL), Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania  
3 Center for Biomedical Image Computing and Analytics (CBICA), Perelman School of Medicine, University of Pennsylvania

✉ Correspondence: [Fengling Hu <[fengling.hu@pennmedicine.upenn.edu](mailto:fengling.hu@pennmedicine.upenn.edu); 423 Guardian Dr, Philadelphia, PA 19104>](mailto:fengling.hu@pennmedicine.upenn.edu; 423 Guardian Dr, Philadelphia, PA 19104)

# CRediT author statement

**Fengling Hu:** Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualization. **Jiayi Tong:** Methodology, Validation, Resources, Investigation, Writing – Review & Editing. **Yong Chen:** Methodology, Validation, Resources, Investigation, Writing – Review & Editing, Supervision. **Russell T. Shinohara:** Conceptualization, Methodology, Validation, Resources, Investigation, Writing – Review & Editing, Supervision, Project administration, Funding acquisition.

# Disclosures and Conflicts of Interest

TODO

# Funding

TODO

# 1 Hello bookdown

All chapters start with a first-level heading followed by your chapter title, like the line above. There should be only one first-level heading (#) per .Rmd file.

## 1.1 A section

All chapter sections start with a second-level (##) or higher heading followed by your section title, like the sections above and below here. You can have as many as you want within a chapter.

### An unnumbered section

Chapters and sections are numbered by default. To un-number a heading, add a {.unnumbered} or the shorter {-} at the end of the heading, like in this section.

# 2 Methods

We develop a distributed GAMLSS algorithm (dGAMLSS) based on the Rigby and Stasinopoulos (RS) algorithm for GAMLSS fitting ([**rigbySemiparametricAdditiveModel1996?**](#ref-rigbySemiparametricAdditiveModel1996); [**rigbyGeneralizedAdditiveModels2005?**](#ref-rigbyGeneralizedAdditiveModels2005)). Ultimately, the dGAMLSS algorithm provides distributed machinery for exact fitting of GAMLSS for distributions with up to four parameters across multiple sites without sharing any patient-level data. For each parameter, dGAMLSS can fit fixed effect covariates as well as up to one smooth term defined as either a fixed effect B-spline or a penalized B-spline with known penalty weight ([**eilersFlexibleSmoothingBsplines1996?**](#ref-eilersFlexibleSmoothingBsplines1996)). Additionally, dGAMLSS enables distributed inference using observed Hessian matrices from each site. Finally, dGAMLSS provides model-based centiles, allowing for normative outcome modeling.

## 2.1 Distributed RS algorithm

The gold-standard, pooled RS algorithm is provided in Appendix B of ([**rigbyGeneralizedAdditiveModels2005?**](#ref-rigbyGeneralizedAdditiveModels2005)). Briefly, the RS algorithm consists of two cycles which result in maximization of the pooled likelihood with respect to the fixed effect and smoothing coefficients. The outer cycle iterates across the GAMLSS distribution-specific parameters in order from to . Meanwhile, the inner cycle performs Newton-Raphson updates on the outer-cycle parameter-wise coefficients while keeping coefficients for all other parameters constant. Each inner cycle is completed when the parameter-wise coefficients converge. The outer cycle is completed when all parameter-wise coefficients converge.

Notably, the pooled RS algorithm describes an additional backfitting cycle within the inner cycle. This backfitting lets GAMLSS fit multiple smooth terms for each parameter, similarly to a generalized additive model (GAM) [TODO]. We forego discussion of the RS backfitting algorithm here – in the distributed setting, backfitting tends to require a prohibitive number of additional communication rounds within each inner cycle. Instead, we provide a distributed RS approach for fitting up to one smooth term per parameter, using either a fixed effect spline design matrix or a penalized spline with a fixed penalty.

We provide some notation:

Total number of sites. Index of sites, .

Total number of subjects across all sites. Number of subjects in site . Index of subjects within a specific site, .

Total number of GAMLSS parameters depending on distribution, , or . corresponds to distributions with , , , and parameters. GAMLSS parameter index, . Pooled fixed effect coefficients for parameter . (Optional) pooled smooth effect coefficients for parameter .

Fitted GAMLSS canonical parameter for site for parameter . , , , and correspond to , , , and , respectively. GAMLSS linear predictor from site for parameter , such that , where is the distribution-dependent canonical link function for the th parameter. Vector of first derivatives of subject-wise log likelihoods with respect to . Diagonal matrix of second derivatives of subject-wise log likelihoods with respect to . Adjusted dependent variable from site for parameter . .

Outcome vector from site , dimension . Fixed effects design matrix from site for parameter , dimension . (Optional) smooth effect design matrix from site for parameter , dimension . Update matrix from site for parameter with dimension Update matrix from site for parameter with dimension

(Optional) Smoothing penalty matrix associated with Fixed smoothing penalty hyperparameter for parameter . Higher corresponds to more smoothing.

Additionally, let indicate row-wise concatenation of matrices.

The abbreviated RS algorithm is as follows:

My alg (??) My LyX table (??)

## 2.2 Distributed RS algorithm

for settings where patient-level data is distributed across multiple centers and direct transfer of the patient-level data is not allowed.